Joint segmentation of multiple aCGH profiles

F. Picard*, E. Lebarbier† and S. Robin†

*UMR CNRS-8071/INRA-1152, Statistique et Génome, Évry.

†AgroParisTech / INRA, UMR518 Unité Mathématiques et Informatique Appliquées, F-75005 Paris.

picard@genopole.cnrs.fr

Segmentation methods have been successfully applied to the mapping of chromosomal abnormalities when using CGH microarrays. Current methods can deal with one CGH profile only, and do not integrate multiple arrays, whereas the CGH microarray technology becomes widely used to characterize chromosomal defaults at the cohort level. In this work, we propose a new statistical model to jointly segment multiple CGH profiles. The basics of our approach is to use mixed linear models to introduce correlations among profiles. We also solve a computational issue which is linked to the use of dynamic programming on heavy datasets. Overall, this procedure offers a statistical and a computational framework for the joint analysis of multiple CGH profiles.

Segmentation models

Segmentation models have shown good performance in the analysis of array CGH data [1]. The objective of such models is to partition the data into segments with homogeneous mean. We model the observed log-ratios by a random process $\{Y_{\ell t}\}_t$ whose mean is subject to $K_\ell - 1$ abrupt changes at breakpoints $\{t_k^\ell\}$ for patient ℓ, (with convention $t_0^\ell = 0$ and $t_{K_\ell}^\ell = n$) and is constant between two breakpoints within the interval $I_k^\ell =]t_{k-1}^\ell, t_k^\ell]$. In this context, the segmentation model is

$$\forall t \in I_k^\ell,\ Y_{\ell t} = \mu_{\ell k} + \varepsilon_{\ell t} \text{ with } \varepsilon_{\ell t}. \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

Interestingly, this model can be put in the general framework of linear models. Considering matrix $\mathbf{T}_\ell$ of dimension $n \times K_\ell$, such that $\mathbf{T}_\ell[t_k^\ell, k] = 1$ and 0 otherwise, and matrix $\boldsymbol{\mu}_\ell = [\mu_{\ell 1}, \dots, \mu_{\ell K_\ell}]$, model (1) can be written as: $\mathbf{Y}_\ell = \mathbf{T}_\ell \boldsymbol{\mu}_\ell + \mathbf{E}_\ell$. More generally, if considering L profiles, we get the following linear model

$$\boxed{\mathbf{Y} = \mathbf{T}\boldsymbol{\mu} + \mathbf{E},}$$

with $\mathbf{E}$ a Gaussian noise with variance $\sigma^2\mathbf{I}$. The particularity of this model is that matrix $\mathbf{T}$ (breakpoints) is estimated, whereas it is fixed in the classical case.

Joint segmentation using Mixed Linear Models

Our hypothesis is that there exists correlation among different profiles at a given position. To model this dependency, we introduce a random effect U_t indenpendent of $\varepsilon_{\ell t}$, and such that $U_t \sim \mathcal{N}(0, \lambda^2)$. Then the model becomes:

$$\forall t \in I_k^\ell,\ Y_{\ell t} = \mu_{\ell k} + U_t + \varepsilon_{\ell t}. \tag{2}$$

In this context, we have $\mathbb{V}(Y_{\ell t}) = \sigma^2 + \lambda^2$, and $cov(Y_{\ell t}, Y_{\ell' t}) = \lambda^2$. Similarly to model (1), model (2) can be written such as:

$$\boxed{\mathbf{Y} = \mathbf{T}\boldsymbol{\mu} + \mathbf{Z}\mathbf{U} + \mathbf{E},}$$

where $\mathbf{Z}$ is the fixed incidence matrix of the random effect with size $[N \times n]$, $N = Ln$ being the total number of data points.

Parameter estimation using the ECM algorithm

We propose to estimate the parameters of the model by maximum likelihood, with $\phi = (\boldsymbol{\mu}, \lambda^2, \sigma^2, \mathbf{T})$ the set of parameters to be estimated. The use of the EM algorithm is now well established in the context of parameter estimation for mixed linear models [2], since those models are incomplete-data model, with $\mathbf{U}$ being the unobserved data. The use of EM lies in the decomposition of the complete-data log-likelihood such that:

$$\log \mathcal{L}(\mathbf{Y}, \mathbf{U}; \phi) = \log \mathcal{L}(\mathbf{Y}|\mathbf{U}; \mathbf{T}, \boldsymbol{\mu}, \sigma^2) + \log \mathcal{L}(\mathbf{U}; \lambda^2).$$

The conditional expectation $Q(\phi; \phi^{(h)})$ of $\log \mathcal{L}(\mathbf{Y}, \mathbf{U}; \phi)$ given $\mathbf{Y}$ is also a sum of two terms $Q_0(\phi; \phi^{(h)})$ and $Q_1(\phi; \phi^{(h)})$ such that:

$$\begin{aligned} -2Q_0(\phi; \phi^{(h)}) &= N\log(2\pi) + N\log\sigma^2 + \|\mathbf{Y} - \mathbf{T}\boldsymbol{\mu} - \mathbf{Z}\widehat{\mathbf{U}}\|^2/\sigma^2 + \text{Tr}\left(\mathbf{Z}\mathbf{W}\mathbf{Z}'\right)/\sigma^2, \\ -2Q_1(\phi; \phi^{(h)}) &= L\log(2\pi) + L\log\lambda^2 + \widehat{\mathbf{U}}'\widehat{\mathbf{U}}/\lambda^2 + \text{Tr}\left(\mathbf{W}\right)/\lambda^2, \end{aligned}$$

where $\widehat{\mathbf{U}} = \mathbb{E}_{\phi^{(h)}}\{\mathbf{U}|\mathbf{Y}\}$ and where $\mathbf{W} = \mathbb{V}_{\phi^{(h)}}\{\mathbf{U}|\mathbf{Y}\}$.

E-step It consists in the calculation of $Q(\phi; \phi^{(h)})$ which only requires the calculation of $\widehat{\mathbf{U}}$ and $\mathbf{W}$. The BLUP is such that $\widehat{\mathbf{U}} = \lambda\mathbf{Z}'\mathbf{V}(\mathbf{Y})^{-1}(\mathbf{Y} - \mathbf{T}\boldsymbol{\mu})$, and we use Henderson's trick which avoids the inversion of $\mathbf{V}(\mathbf{Y})$. So we get at iteration $(h+1)$

$$\widehat{\mathbf{U}}^{(h+1)} = \mathbf{W}^{(h)}\mathbf{Z}'\left(\mathbf{Y} - \mathbf{T}^{(h)}\boldsymbol{\mu}^{(h)}\right)/\sigma^{2(h)}, \quad \mathbf{W}^{(h+1)} = \sigma^{2(h)}\left(\mathbf{Z}'\mathbf{Z} + \frac{\sigma^{2(h)}}{\lambda^{2(h)}}\mathbf{I}\right)^{-1}.$$

CM-steps The principle of the ECM algorithm [3] is to breakdown the maximization of $Q(\phi; \phi^{(h)})$ with respect to ϕ (global M-step) into simpler CM-steps which focus on one parameter, the others being fixed. Explicit formulas exist for $\left\{\lambda^{2(h+1)}, \sigma^{2(h+1)}\right\}$, and the challenging step is the update of the breakpoints:

$$\left\{\mathbf{T}^{(h+1)}, \boldsymbol{\mu}^{(h+1)}\right\} = \arg\max_{\mathbf{T}, \boldsymbol{\mu}} Q_0\left(\phi; \lambda^{2(h+1)}, \sigma^{2(h+1)}\right).$$

This optimization problem is equivalent to the minimization of the residual sum of squares:

$$SSR_K(\boldsymbol{\mu}, \mathbf{T}) = \|\mathbf{Y} - \mathbf{T}\boldsymbol{\mu} - \mathbf{Z}\widehat{\mathbf{U}}^{(h+1)}\|^2/\sigma^{2(h+1)} = \sum_{\ell=1}^{L}\sum_{k=1}^{K_\ell} SSR_k^\ell(\boldsymbol{\mu}_\ell, \mathbf{T}_\ell) \tag{3}$$

under the constraint $\sum_\ell K_\ell = K$. This sum is additive according to the number of segments, which allows us to use the dynamic programming algorithm.

Dynamic programming on heavy datasets

Dynamic programming is an efficient method to estimate breakpoints when the number of segments K is given. This algorithm can be used when the function to be optimized is additive with respect to the number of segments, such that:

$$SSR_{K_\ell}^\ell(\boldsymbol{\mu}_\ell, \mathbf{T}_\ell) = \sum_{k=1}^{K_\ell}\sum_{t \in I_k^\ell}(Y_{\ell t} - \hat{\mu}_{\ell k}),$$

and its complexity is $\mathcal{O}(n^2)$. However using Dynamic Programming may be impossible if n is large, and especially when dealing with multiple CGH profiles, the number of points is $L \times n$. When $SSR_K(\boldsymbol{\mu}, \mathbf{T})$ can be written as in Equation (3), we propose to reduce the complexity of the segmentation step using a 2-stage Dynamic Programming.

Stage-1. We denote by $SSR_k^\ell(J^\ell)$ the residual sum of squares when segmenting profile J^ℓ into k segments. This segmentation step is based on the calculus of $SSR_1^\ell(]i, j])$ and on the recursive minimization

$$\forall k \in [1 : K_\ell],\ SSR_k^\ell(]t_1^\ell; j]) = \min_h \left\{SSR_{k-1}^\ell(]t_1^\ell, h]) + SSR_1^\ell(]h, j])\right\}.$$

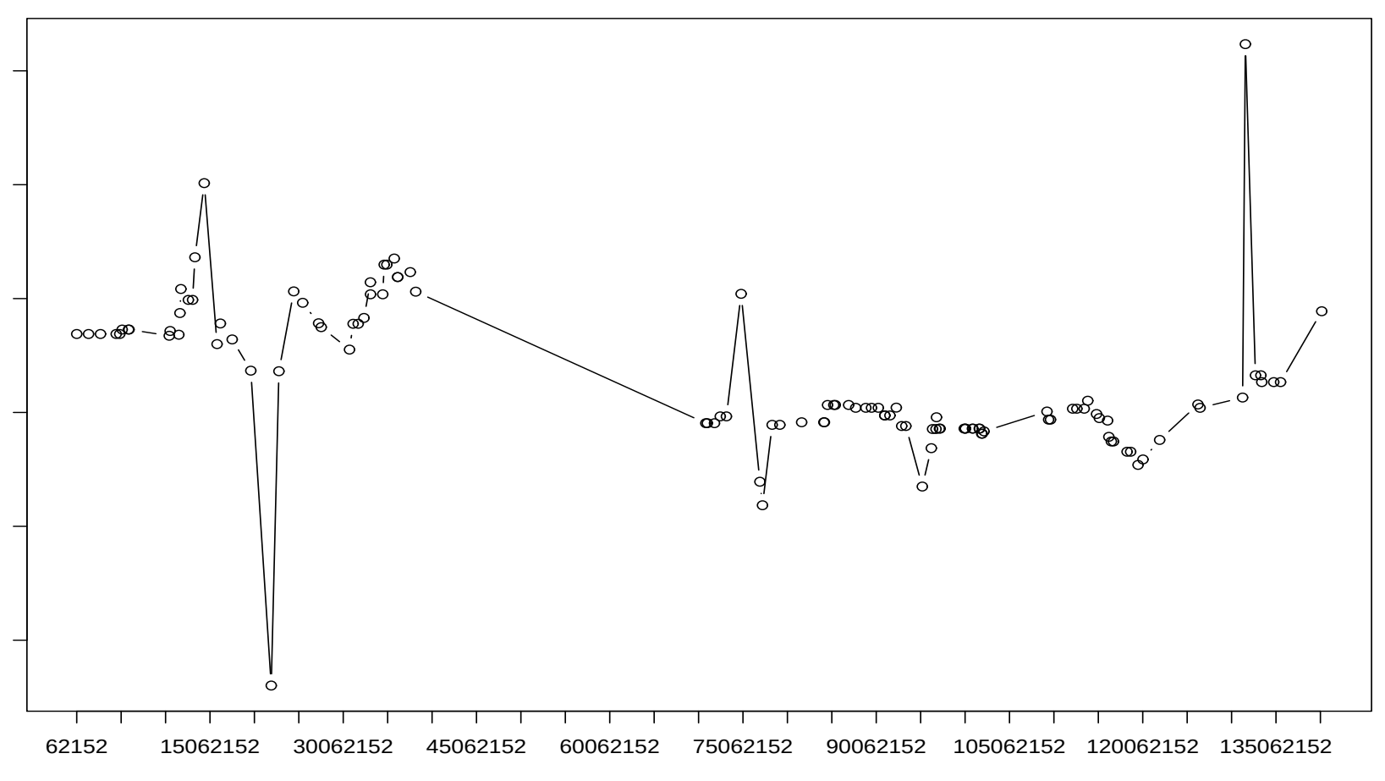
Stage-2. It consists in the repartition of segments among patients. We denote by $SSR_K(J^1, \dots, J^\ell)$ the total sum of squares for a model with K segments spread among ℓ patients. This step is based on the calculus of $SSR_k^\ell(J^\ell)$ which has been done in Step-1, and on the recursive minimization:

$$\forall \ell \in [1 : L],\ SSR_K(J^1, \dots, J^\ell) = \min_{k'+k''=K}\left\{SSR_{k'}(J^1, \dots, J^{\ell-1}) + SSR_{k''}^\ell(J^\ell)\right\}.$$

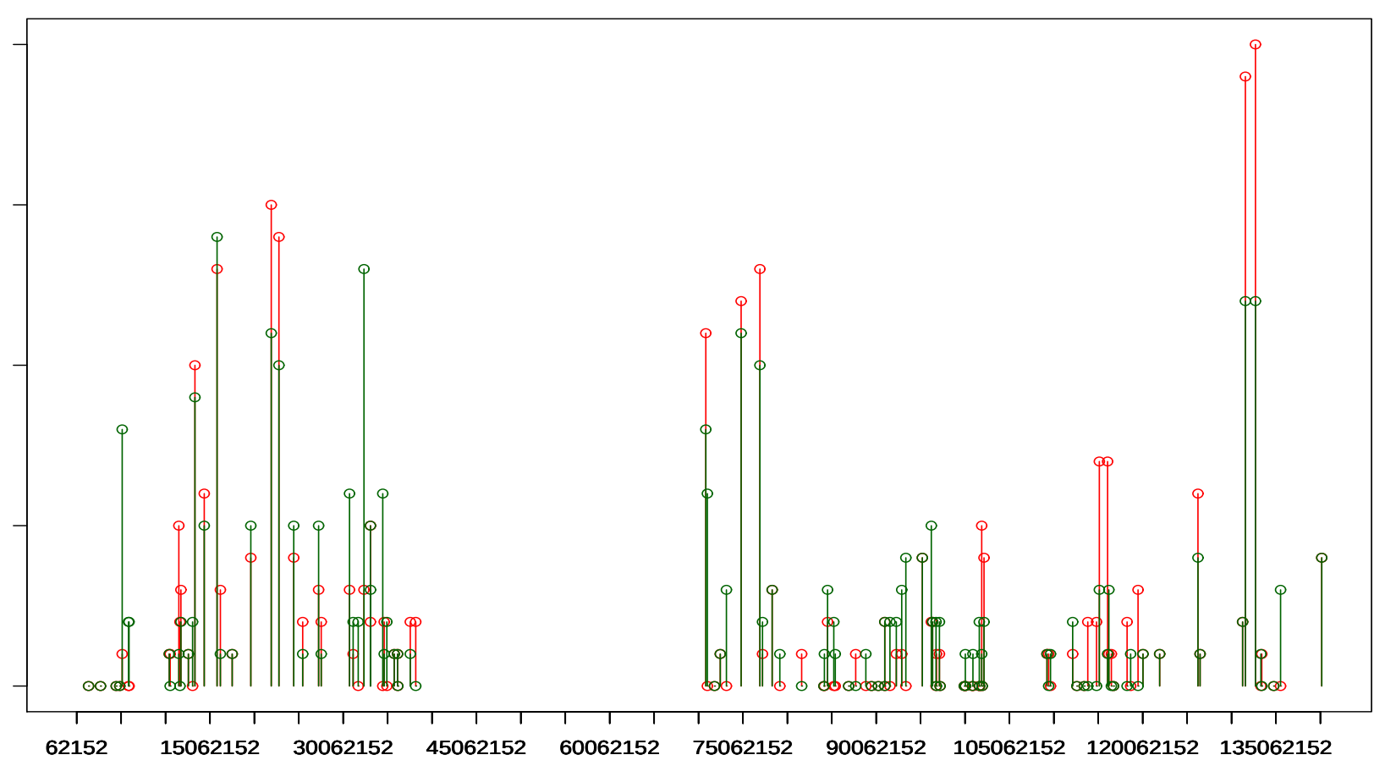
Complexity If all series have the same length n (so $N = Ln$) and are segmented into $K_\ell = k$ segments each (so $K = Lk$) and assuming that $k = \alpha n$ (with $\alpha \ll 1$), the two-stage dynamic programming algorithm has a complexity of $\mathcal{O}(\alpha L n^2[n + \alpha L^2])$ whereas the overall one has complexity $\mathcal{O}(\alpha L^3 n^3)$.

Application

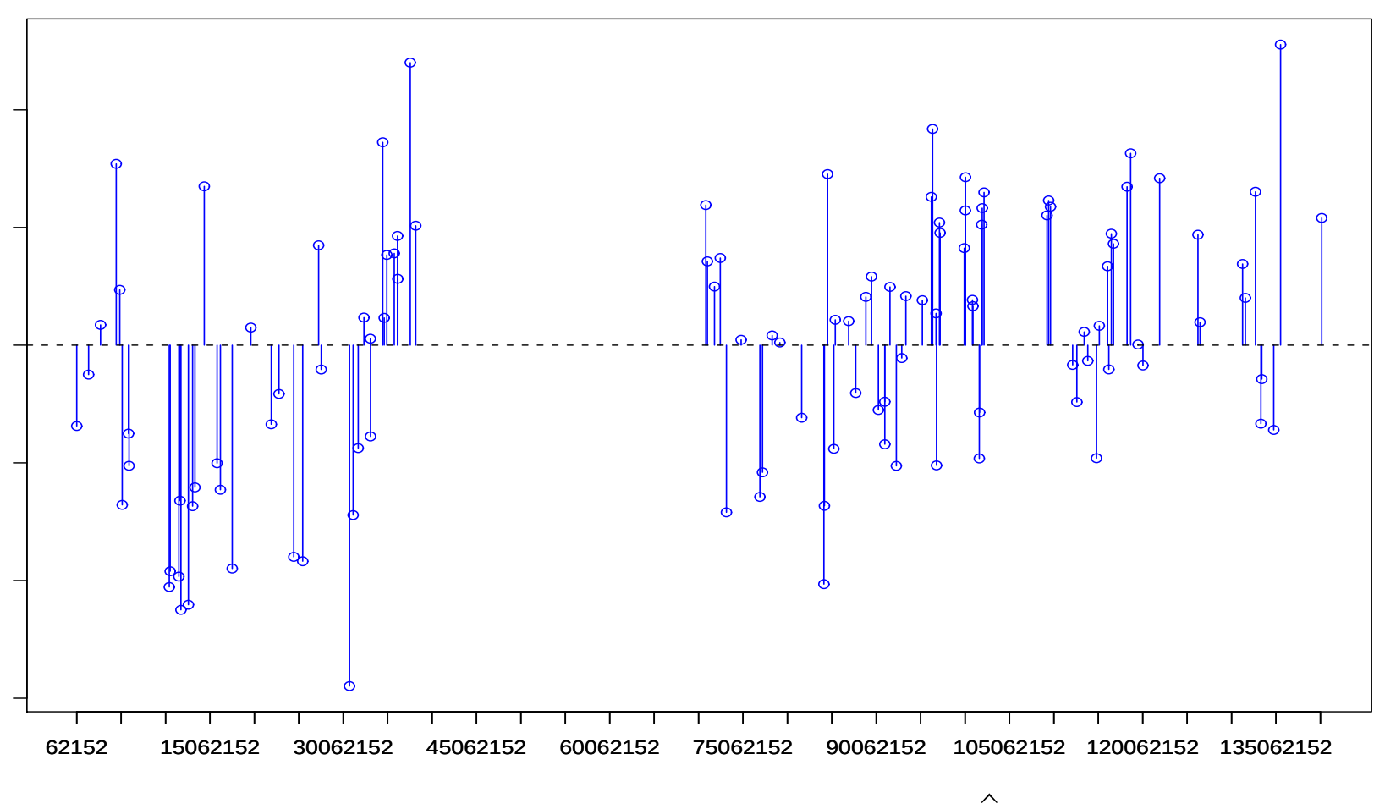
The data consists in a series of 57 bladder tumors, chromosome 9, described in [4]. The total number of segments is $\hat{K} = \sum_\ell \hat{K}_\ell$ where $\hat{K}_\ell$ has been estimated using **CGH_Seg** [1]. This allows us to focus on the breakpoint positions only. When looking at parameter estimates, $\sigma^2 = 3.92\,10^{-3}$ and $\lambda^2 = 3.12\,10^{-3}$ which leads to a correlation of 0.44 among positions. As a result, breakpoint positions change between methods. Using the joint segmentation model, we identify more breakpoints at BACs CTB-65D18 (p-arm) and RP11-14J9 (q-arm), which contain proteins known to be involved in bladder tumor, like protein p16. An interesting feature is also the enrichment in breakpoints near the telomere. This result will have to be further investigated. Finally, we will have to interpret the values of $\hat{\mathbf{U}}$. High values of this effect could indicate common characteristics of BACs that may be caused by technical artifacts.



Mean CGH profile among 57 patients



Breakpoints frequency.
— joint segmentation,— individual segmentation



Predicted random effect $\hat{\mathbf{U}}$

References

[1] Picard et al. (2005), A statistical approach for array CGH data analysis, *BMC Bioinformatics*,6(27).

[2] van Dyck (2000), Fitting mixed-effects models using efficient EM-type algorithms, *Jour. Comp. and Graph. Statistics*, 9, 78–98.

[3] X.-L. Meng and D.B. Rubin (1993), Maximum likelihood estimation via the ECM algorithm: a general framework, *Biometrika*,80(2), 267–278.

[4] Stransky et al. (2006), Regional copy number-independent deregulation of transcription in cancer, *Nat. Gen.*,38(12), 1386–96.